1 The Markov inequality and Chernov bounds

Proposition 1.1 (Markov) Let $Y$ be a positive random variable. Then for all $t > 0$,

$$
P(Y \geq t) \leq \frac{E(Y)}{t}.
$$

Proof: The following uses the pdf $f_Y(y)$ for illustrative purposes, but this can also be argued in a similar way if the density does not exist. We have

$$
E(Y) = \int_0^\infty y f_Y(y) \, dy
$$

$$
\geq \int_t^\infty y f_Y(y) \, dy
$$

$$
\geq t \int_t^\infty f_Y(y) \, dy
$$

$$
= t \cdot P(Y \geq t)
$$

Corollary 1.2 Let $Y$ be a positive random variable. Then for all $t > 0$,

$$
P(Y \geq t) \leq \frac{E(e^{\lambda Y})}{e^{\lambda t}}, \quad \text{for any } \lambda > 0.
$$

Proof: This follows from the Markov inequality and the fact that

$$
P(\phi(Y) \geq \phi(t)) = P(Y \geq t)
$$

for any $\phi(\cdot)$ that is strictly monotonically increasing.

Corollary 1.3 (Chernov) Let $X_1, X_2, \ldots, X_M$ be iid random variables. Then

$$
P\left( \sum_{m=1}^M X_m \geq t \right) \leq \frac{(E(e^{\lambda X_1}))^M}{e^{\lambda t}}.
$$

Proof: Apply the previous corollary with

$$
E\left(e^{\lambda(X_1+X_2+\cdots+X_M)}\right) = E(e^{\lambda X_1}) \cdot E(e^{\lambda X_2}) \cdots E(e^{\lambda X_M}) = (E(e^{\lambda X_1}))^M
$$

2 Gaussian concentration

Proposition 2.1 Let $u \in \mathbb{R}^N$ be a fixed vector with $\|u\|_2 = 1$, and let $\Phi$ be an $M \times N$ Gaussian random matrix with iid entries

$$\Phi_{m,n} \sim \text{Normal}(0, M^{-1}).$$

Set $v = \Phi u$ and fix $0 < \delta < 1$. Then

$$\mathbb{P}(\|v\|_2^2 \geq 1 + \delta) \leq e^{-(\delta^2 - \delta^3)M/4}.$$

Proof. The entries $v_n$ of $v$ are independent Gaussian random variables with mean 0 and variance $\|u\|_2^2/M = 1/M$. Using the Chernov bound, we have

$$\mathbb{P}(\|v\|_2^2 > 1 + \delta) \leq \frac{\mathbb{E}(e^{\lambda v_1^2})^M}{e^{\lambda(1+\delta)}}.$$

We know that $v_1^2$ is a chi-squared random variable, and $\mathbb{E}(e^{\lambda v_1^2})$ is its moment generating function. We can look up the mgf on Wikipedia\(^1\),

$$\mathbb{E}(e^{\lambda v_1^2}) = \frac{1}{\sqrt{1 - 2\lambda/M}}, \quad \lambda < M/2.$$

Combining this with the above, we have

$$\mathbb{P}(\|v\|_2^2 > 1 + \delta) \leq \left(\frac{e^{-2\lambda(1+\delta)/M}}{1 - 2\lambda/M}\right)^{M/2}, \quad \text{for all } \lambda < M/2.$$

We choose the particular value of

$$\lambda = \frac{M\delta}{2(1+\delta)} < M/2$$

that makes the right hand side as small as possible, so we now have

$$\mathbb{P}(\|v\|_2^2 > 1 + \delta) \leq ((1+\delta)e^{-\delta})^{M/2}. \quad (1)$$

It is a fact that

$$1 + \delta \leq e^{\delta - (\delta^2 - \delta^3)/2}, \quad \text{for } 0 \leq \delta \leq 1. \quad (2)$$

You can prove this last statement using the Taylor theorem or by simply plotting the two functions in MATLAB:

\(^1\)https://en.wikipedia.org/wiki/Chi-squared_distribution
The proposition follows by applying (2) to (1).

**Proposition 2.2** Let $u, v, \Phi$ and $\delta$ be as in Proposition 2.1. Then
\[
\mathbb{P}(\|v\|_2^2 \leq 1 - \delta) \leq e^{-(\delta^2 - \delta^3)/4}.
\]

**Proof.** We take
\[
\mathbb{P}(\|v\|_2^2 < 1 - \delta) = \mathbb{P}(\|v\|_2^2 > \delta - 1),
\]
and follow pretty much the same steps as last time:
\[
\mathbb{P}(\|v\|_2^2 > \delta - 1) \leq \left(\frac{e^{-\lambda\|v\|_2^2}}{e^{\lambda(\delta - 1)}}\right)^M, \quad \lambda \geq 0,
\]
\[
\leq \left(\frac{e^{2(1-\delta)\lambda/M}}{1 + 2\lambda/M}\right)^{M/2}, \quad \text{for all } \lambda \geq 0,
\]
\[
= \left((1 - \delta)e^{\delta}\right)^{M/2}, \quad \text{by taking } \lambda = \frac{M\delta}{2(1 - \delta)},
\]
\[
\leq e^{-(\delta^2 - \delta^3)/4}.
\]

Combining the two propositions, we have
\[
\mathbb{P}(\|v\|_2^2 - 1 > \delta) \leq 2e^{-(\delta^2 - \delta^3)/4}.
\]
Note that for $0 < \delta < 1/2$, we have $\delta^2 - \delta^3 > \delta^2/2$ we can simplify this to
\[
\mathbb{P}(\|v\|_2^2 - 1 > \delta) \leq 2e^{-\delta^2/8}.
\]

### 3 The Johnson-Lindenstrauss Lemma

We now have all the pieces we need to prove the JL Lemma.
Lemma 3.1 (Johnson-Lindenstrauss) Let \( Q = \{x_1, \ldots, x_Q\} \) be a set of \( Q = |Q| \) vectors in \( \mathbb{R}^N \). Let \( \Phi \) be an \( M \times N \) Gaussian random matrix whose entries are iid with mean 0 and variance \( M^{-1} \). Fix a failure probability \( \epsilon \) between zero and one, and a precision \( \delta \) between zero and one. If

\[
M \geq \frac{8}{\delta^2} \left( 2 \log Q + \log(1/\epsilon) \right),
\]

then with probability at least \( 1 - \epsilon \), \( \Phi \) will uniformly preserve all the distances between points in \( Q \),

\[
(1 - \delta) \|x_i - x_j\|^2 \leq \|\Phi x_i - \Phi x_j\|^2 \leq (1 + \delta) \|x_i - x_j\|^2, \quad \text{for all } x_i, x_j \in Q.
\]

Proof. To ease the notation, let \( d_{ij} = \|x_i - x_j\| \). Using the union bound then the Gaussian concentration results above, we have

\[
P \left( \max_{x_i, x_j \in Q} \|\Phi(x_i - x_j)\|^2 - d_{ij}^2 > \delta d_{ij}^2 \right) \leq \sum_{x_i, x_j \in Q} P \left( \|\Phi(x_i - x_j)\|^2 - d_{ij}^2 > \delta d_{ij}^2 \right) \leq \frac{Q^2}{2} e^{-\delta^2 M / 8} \leq Q^2 e^{-\delta^2 M / 8}.
\]

We can make the right-hand side above less than \( \epsilon \) by taking \( M \) as in (4).

\[ \blacksquare \]

4 Covering and packing numbers

A \( \gamma \)-net \( \mathcal{N}_\gamma \) for a compact set \( S \subset \mathbb{R}^N \) is a finite set of points \( u_1, \ldots, u_Q \) such that

\[
\min_{u \in \mathcal{N}_\gamma} \|u - x\|_2 \leq \gamma \quad \text{for all } x \in S.
\]

The covering number \( \mathcal{N}_\gamma(S) \) for \( S \) is the size of the smallest possible \( \gamma \)-net. If \( B_k \) is the unit ball in \( \mathbb{R}^k \),

\[
B_k = \{x \in \mathbb{R}^k : \|x\|_2 \leq 1\},
\]

then it is a classical result from analysis that

\[
\mathcal{N}_\gamma(B_k) \leq \left(1 + \frac{2}{\gamma}\right)^k. \tag{5}
\]

The following\(^2\) shows us how we can approximate the supremum of quadratic functionals over the sphere with the maximum over a \( \gamma \)-net.

Proposition 4.1 Let \( H \) be a symmetric \( N \times N \) matrix, and let \( \mathcal{N}_\gamma \) be an \( \gamma \)-net for the unit sphere \( B_N \). Then

\[
\|H\| = \sup_{x \in B_N} |x^T H x| \leq \left(\frac{1}{1 - 2\gamma}\right) \cdot \max_{x \in \mathcal{N}_\gamma} |x^T H x|.
\]

\(^2\)This appears as Lemma 5.4 in Roman Vershynin’s excellent manuscript “Introduction to the non-asymptotic theory of random matrices”.

4
Proof. Let $v$ be the eigenvector of $H$ corresponding to its largest-magnitude eigenvalue; then this vector achieves the supremum on the left above, $|v^T H v| = \|H\|$. Let $u \in \mathcal{N}_\gamma$ be within $\gamma$ of $v$, $\|u - v\|_2 \leq \gamma$. Then
\[
|v^T H v| - |u^T H u| \leq |v^T H v - u^T H u| = |(v - u)^T H v + u^T H (v - u)| \\ \leq \|H\| \|v - u\|_2 \|v\|_2 + \|H\| \|v - u\|_2 \|u\|_2 \\ \leq 2\gamma \|H\|.
\]
Thus for this particular $u$,
\[
|u^T H u| \geq (1 + 2\gamma) \|H\|,
\]
and so the same must be true for the maximum over the entire net $\mathcal{N}_\gamma$. 

5 Subspace embedding

Proposition 5.1 Let $S$ be a $K$-dimensional subspace of $\mathbb{R}^N$, and let $\Phi$ be an $M \times N$ Gaussian random matrix whose entries are iid with mean 0 and variance $M^{-1}$. Fix a failure probability $\epsilon$ between zero and one, and a precision $\delta$ between zero and one. If
\[
M \geq \frac{10}{\delta^2} (3.72 K + \log(2/\epsilon)),
\]
then with probability at least $1 - \epsilon$, $\Phi$ will uniformly preserve all the distances between points in $S$,
\[
(1 - \delta) \|x_1 - x_2\|_2^2 \leq \|\Phi x_1 - \Phi x_2\|_2^2 \leq (1 + \delta) \|x_1 - x_2\|_2^2, \quad \text{for all } x_1, x_2 \in S. \tag{6}
\]

Proof. It is enough to show that
\[
\sup_{x \in S, \|x\|_2 = 1} \|\Phi x\|_2^2 - 1 \leq \delta
\]
with probability at least $1 - \epsilon$. Taking $\gamma = 1/20$, it is enough to show that
\[
\max_{x \in \mathcal{N}_{1/20}} \|\Phi x\|_2^2 - 1 \leq 9\delta/10. \tag{7}
\]
Combining our Gaussian concentration result (3) along with the covering number estimate (5), we can take a union bound just like in the proof of JL:
\[
P((7) \text{ holding}) \leq 2(41)^K \epsilon^{-81M\delta^2/100}
\]
We can make this less than $\epsilon$ for
\[
M \geq \frac{10}{\delta^2} (3.72 K + \log(2/\epsilon)).
\]
\[\square\]
6 Compressed least squares

The following is such an immediate result of the above that we will not even state it as a proposition.

Let \( A \) be an \( M \times N \) matrix with \( M \gg N \); we will assume that \( A \) has full column rank. Let \( y \in \mathbb{R}^M \) be a fixed vector, and let \( \hat{x} \in \mathbb{R}^N \) be the least squares solution

\[
\hat{x} = \arg \min_{x \in \mathbb{R}^N} \| y - Ax \|_2^2.
\]

Suppose that \( \Phi \) obeys (6) for \( S = \text{Range}(\begin{bmatrix} A & y \end{bmatrix}) \), the (at most) \( K + 1 \) dimensional subspace spanned by the columns of \( A \) and \( y \). Then the compressed least-squares solution

\[
\hat{x} = \arg \min_{x \in \mathbb{R}^N} \| \Phi y - \Phi Ax \|_2^2.
\]

Then \( \hat{x} \) is at most \( \delta \)-suboptimal for the original least-squares problem,

\[
\| y - A\hat{x} \|_2^2 \leq (1 + \delta)\| y - A\hat{x} \|_2^2.
\]